Ministry of Transportation
Ministère des Transports
Ontario

91-01652/

Research and Development Br.
Central Bldg., Room 331
1201 Wilson Ave.
Downsview, Ontario M3M 1J8

Tel.: (416) 235-4707
Fax.: (416) 235-4872

Re: R & D Report TDS-90-04: Blackspot Identification and Treatment Evaluation (R & D Project 25197)

The attached report describes a new method for the estimation of accident potential for Ontario roads. The method first uses regressions to produce initial estimates of accident potential on the basis of traffic and geometric characteristics. These estimates are then combined with accident counts, using an empirical Bayesian procedure, to produce final estimates of accident potential. The empirical Bayesian estimates are shown to be superior to estimates based on the accident count only, particularly for short, high-accident sections that may be of interest in a blackspot identification and treatment program.

The report was prepared by Professor Bhagwant N. Persaud (Ryerson Polytechnical Institute, Toronto) as part of a comprehensive research program under the overview of the Roadway Safety Committee of Provincial Highways Program.

We hope that you will find this research useful. The proposed method of estimating accident potential is recommended for use in the identification of accident blackspots and in evaluation of safety treatments in the Province of Ontario.

Yours sincerely,

Peter Smith
Director
Research & Development Br.

Encl.

Ontario
Ministry of Transportation
Research and Development Branch

Documentation Page

TDS-90-04

Blackspot Identification and Treatment Evaluation

Author(s): B. N. Persaud -Department of Civil Engineering, Ryerson Polytechnical Institute

Date Published: November, 1990

Published by: Research and Development Branch, MTO

Contact Person: Alex Ugge, Pavements and Roadway Office
(416) 235-4698

Abstract: A method for the estimation of accident potential for Ontario roads is developed. The method first uses regression models to produce an initial estimate of accident potential on the basis of a section's traffic and geometric characteristics. This estimate is then combined with the section's accident count, using an empirical Bayesian procedure, to produce final estimates of the section's accident potential. The empirical Bayesian estimates are shown to be superior to estimates based on the accident count only, particularly for short, high-accident sections that might be of interest in a blackspot identification and treatment program. The proposed method of estimating accident potential is recommended for use in the identification of accident blackspots and in evaluation of safety treatments.

Comments: This report is a part of Research and Development Project 25197

Key Words: Blackspots, accident analysis, regression models, accident potential, empirical Bayesian, traffic safety, GLIM

Distribution: not restricted

Copyright Status: Crown copyright © 1990 Ministry of Transportation

Blackspot Identification and Treatment Evaluation

B. N. Persaud
Department of Civil Engineering
Ryerson Polytechnical Institute

Published by
The Research and Development Branch
Ontario Ministry of Transportation
Hon. E. Philip, Minister
P. Jacobsen, Deputy Minister

Published without prejudice as to the application of the findings.
Crown copyright reserved; however, this document may be reproduced for non-commercial purposes with attribution to the Ministry.

For additional copies, contact:
The Editor, Technical Publications
Room 320, Central Building
1201 Wilson Avenue
Downsview, Ontario
Canada M3M 1J8

Telephone: (416) 235-3480
Fax: (416) 235-4872

November 1990

Table of Contents

1/ Introduction **1**

2/ Data **4**

2.1/ Raw Data 4

2.2/ Final Datasets 4

3/ Procedures, Methods and Theoretical Foundations **6**

3.1/ Regression Models for Initial Prediction 7

3.2/ Empirical Bayesian Revised Estimates 8

3.3/ Illustrating the Empirical Bayesian Method 8

3.4/ Further Exploration of the Methodology 9

4/ Analysis and Results **11**

4.1/ Regression Models for Estimating Accident Potential 11

4.2/ Empirical Bayesian Estimation and Validation 19

5/ Summary, Conclusions and Recommendations **24**

5.1/ Summary and Conclusions 24

5.2/ Recommendations 25

References **27**

Appendices:

A/ Updating Empirical Bayesian Estimates 28

B/ Illustrative Applications 31

List of Tables and Figures

Tables

1/ Summary information for final datasets 5
2/ Level 1 regression models 12
3/ Categories for class 2 roads 14
4/ Level 2 regression models for class 2 roads 15
5/ Categories for class 3 roads 17
6/ Level 2 regression models for class 3 roads 18
7/ Accident potential of selected rural 2-lane class 2 roads 21

Figures

1/ Illustrating the procedure for estimating accident potential 6
2/ Regression models for class 2 roads (1km) 16
3/ Regression models for class 3 roads (1km) 19
A1/ Gamma distributions of accident potential 30

1/ Introduction

The efficient allocation of resources to highway safety treatments requires that accident blackspots (unsafe intersections, road sections, etc.) be properly identified and that the estimates of safety benefit of potential treatment be as sound as possible. Crucial to both aspects is the sound estimation of accident potential of a site under consideration for treatment.

In the case of safety benefits, the safety effect of a treatment is usually estimated from it's previous applications by comparing the accident potential of sites before and after they were treated. This estimate is then applied to the current estimate of accident potential of a site under consideration for the treatment.

In the case of blackspot identification, accident potential is often used as an initial screening mechanism to quickly identify a manageable number of possible blackspot sites for a second stage -- a close examination of all relevant data for the identified sites, usually supplemented by field inspection. Because this second stage is time-consuming and costly, and because a very large number of possible sites come under the jurisdiction of organizations such as the MTO, it is important that the initial blackspot screening mechanism be as efficient as possible.

The common methods of identifying possible blackspots may not only be inefficient but can also cause the safety effect of treatments to be overestimated. Many jurisdictions use variants of the "rate-and-number" method (1) in which sites registering an unusually high accident count and/or accident rate (per vehicle-km, sometimes weighted by severity) are selected for inspection. The efficiency of these methods is limited by the well-known fact that accident counts at a site fluctuate from year to year. If the number of accidents in the past 1-3 years on a relatively safe site is randomly high, the site can be identified as a possible blackspot and be subjected to detailed inspection, possibly unnecessarily.

To compound this problem, if such a site is treated, comparison of the randomly high accident count before treatment with what happened after will almost surely overestimate the safety effect of the treatment. When this overestimate is later used in estimating the cost effectiveness of applying the treatment at another site, the result could be a gross overestimate of safety benefits if that site, too, had a randomly high accident count. At the other extreme, a truly hazardous site can escape detection, or the safety effect of treating it can be underestimated, because of a random down-fluctuation in the accident count at that site.

These difficulties can be illustrated by considering a resource allocation procedure used by the U.S. Federal Railroad Administration (FRA) (**2**) to determine which rail-highway crossings should be upgraded. In the past there was an understandable tendency to upgrade those crossings which had accidents. Studies comparing before-and-after accident counts revealed impressive accident reductions of about 70% for upgrading from cross-bucks to flashers.

It has been shown (**3**) that, because the upgraded crossings generally had randomly high accident counts, the true reduction is around 50%, not 70%, and that the 50% should be applied to the true accident potential of crossings under consideration, not to the short-term count.

To illustrate the impact of these problems on a resource allocation procedure, and using some realistic numbers, suppose a crossing equipped with a crossbuck had 3 accidents in the past 3 years. Using methods outlined in (**3**), the true (long-term) accident potential at this crossing is say 0.2 accidents per year (not the randomly high value of 1 per year). Using a cost of $50,000/accident, four separate estimates of safety benefits are provided below.

Savings Per Year		
	Accidents	**Dollars**
A. Apply 70% to short term count	0.70	$35,000
B. Apply 50% to short term count	0.50	$25,000
C. Apply 70% to long term mean	0.14	$ 7,800
D. Apply 50% to long term mean	0.10	$ 5,000

This illustration brings out the dramatic discrepancy between the correct estimate (D) and the one (A) that would be used in the absence of knowledge on the true safety effect and on the true accident potential. In fact, the result of using either (A) or (B) in a resource allocation procedure, could be a perpetuation of resource allocation to entities with a random up-fluctuation in accident count. This clearly is not appropriate to achieving maximum safety benefits from limited funds, particularly in the light of public pressure to apply safety treatments where the recent accident count was relatively high.

The efficiency of a resource allocation process for safety improvements can be maximized if initial blackspot screening and treatment evaluation be both based on the average number of accidents in the long run on a site (the true accident potential) rather than on the short term count. For obvious practical reasons, this long term accident potential cannot be observed, so has to be estimated.

The primary focus of the research described in this report was to explore how to obtain sound estimates of accident potential for Ontario Road sections and how to use these estimates in blackspot identification and in the evaluation of safety treatment.

2/ Data

2.1/ Raw Data

Raw computerized data was obtained from 2 MTO sources. Inventory information for LHRS (Linear Highway Referencing System) sections in the province was obtained from the Transportation Capital Branch. The record for each section (or sub-section) contained information on whether the highway was divided or undivided, the highway category and environment, section length, lane and shoulder widths, roadway and shoulder surface type, number of lanes, speeds and speed limits, along with other geometric information.

The other raw data set contained, for each of the years 1983 - 1986, the total number of accidents and detailed traffic information for LHRS sections (or sub-sections). The traffic dataset included information on % commercial vehicles, seasonal and annual average daily traffic volumes, and directional split.

2.2/ Final Datasets

Final datasets were prepared for use in the analysis by combining the raw traffic and inventory data for as many LRHS sections or sub-sections as possible.

The main problem encountered was that occasionally an LHRS section would be split sometime during 1983-1986. For modelling purposes, such sections had to be re-combined as one by aggregating the accidents and section lengths. A consistency check on the accident data flagged all sections with a large variance to mean ratio. The vast majority of sections so flagged appeared to be those that were split during the study period. The few remaining sections with inconsistent accident data were deleted from the dataset.

The full combined dataset was broken down into 3 road classes. Summary information for each final dataset is shown in Table 1.

Table 1/ Summary information for final datasets

	CLASS 1 Freeways Highways	CLASS 2 Other Kings Tertiary	CLASS 3 Secondary
Number of sections	404	1680	166
Total length (km)	1583.6	12024.6	2396.5
Average Daily Traffic (Weighted by Length)			
1983	24298	2548	318
1984	25129	2608	326
1985	26219	2610	324
1986	28981	2767	338
Total Accidents			
1983	11612	15853	564
1984	12533	16963	601
1985	14641	18125	662
1986	14420	17033	554

The summary figures in Table 1 show that, for the period 1985-1986, the Class 1 roads experienced a 20% increase in accidents and a 12% increase in traffic over 1983-1984. Comparable figures for the other classes show that Class 2 roads had a 7% increase in accidents and a 4% increase in traffic while Class 3 roads had a 4% increase in accidents and a 3% increase in traffic.

3/ Procedures, Methods and Theoretical Foundations

The procedure for developing and applying the models for estimating accident potential of road sections follows closely on that used for estimating accident potential for Ontario drivers (**4**) and for rail-highway crossings (**3**). Basically, what is desired is an estimate of the accident potential of a road section, given (i) its accident history and (ii) its traffic, geometric and other characteristics. The method used to get this estimate is, in effect, an Empirical Bayesian procedure that mixes the two pieces of information. This procedure is depicted in Figure 1 and is described below as a sequence of two steps, followed by an example application and some further exploration of estimation process.

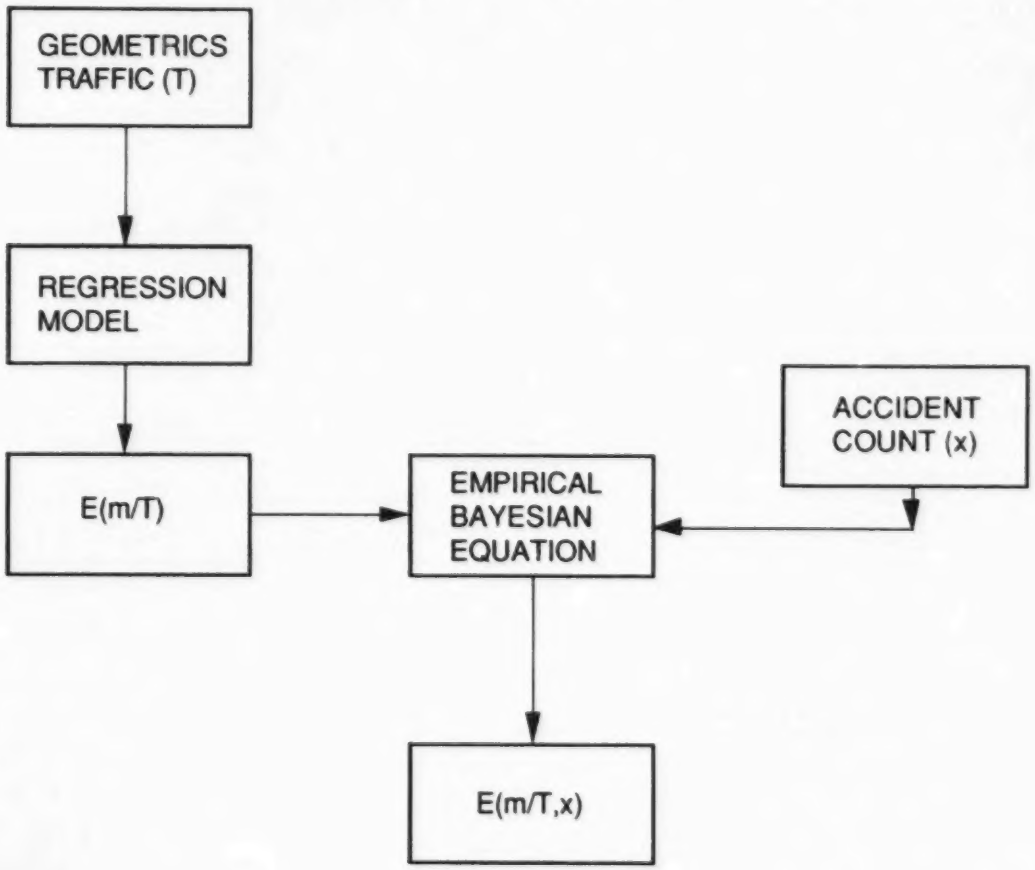

Figure 1/ Illustrating the procedure for estimating accident potential

3.1/ Regression Models for Initial Prediction of Accident Potential (Step 1)

Generalized linear modelling using the GLIM computer package (**5**) was used to estimate E(m|T), the (long term) accident potential of a section for the 2 year period, 1983-1984, given its characteristics, T (e.g., ADT), for this period. A longer period was not used because it is likely that the characteristics of many sections would have changed significantly. The model form used was:

$$E(m|T) = (SCL) \times a1 \times (ADT)^{b1} \qquad (1)$$

where SCL is the section length in km, and a1, b1 are model parameters estimated by GLIM.

The accident count on a section was used as an estimate of the dependent variable, E(m|T). GLIM allows the specification of a negative binomial error structure for the dependent variable which is now known to be more appropriate for accident counts than the traditional normal distribution.

With a negative binomial error specification, it can be shown that the variance of the regression estimates can be estimated from:

$$Var(m|T) = [E(m|T)]^2/k \qquad (2)$$

where k can be estimated using a maximum likelihood procedure that assumes that each squared residual of the regression model is an estimate of Var(m|T) and that each count comes from a negative binomial distribution with mean E(m|T) and variance given by Equation 2.

Two levels of models were attempted for each road class listed in Table 1. Level 1 used ADT as the only independent variable, while the more elaborate Level 2 models used other "geometric" variables. The Level 2 models were so constructed that the parameters a1 and b1 depend on the values of the relevant geometric variables.

3.2/ Empirical Bayesian Revised Estimates of Accident Potential (Step 2)

In general, two road sections which are similar in all of the independent variables used in the regression model (e.g., ADT, section length and road class for the Level 1 models) will still be different in true accident potential even though they will have the same model predictions according to equation 1. This is because it is not possible to account in the model for all the factors that cause differences in accident potential (e.g., weather, geometrics).

To account for this shortfall, $E(m|T)$ from Equation 1 can be further refined for an individual road section using the accident count, x, on that section to give $E(m|T,x)$, a final revised estimate of accident potential. It can be shown that, if the variation in $m|T$ can be described by a gamma probability distribution and if accident occurrence on a section follows the Poisson probability law (both shown to be reasonable assumptions that are consistent with the negative binomial error structure assumed earlier), then the revised estimate of accident potential is:

$$E(m|T,x) = x + [E(m|T)/(E(m|T)+Var(m|T))]*[E(m|T)-x] \qquad (3)$$

and that,

$$Var(m|T,x) = [x + E(m|T)^2/Var(m|T)]/[1 + E(m|T)/Var(m|T)]^2 \qquad (4)$$

The value obtained from Equation 3 is known as an empirical Bayesian estimate. It turns out that the variation in $\{m|T,x\}$ can also be described by a gamma distribution.

Note that in Equations 3 and 4, the observation period for the accident count must be of the same length as that for which the regression prediction $E(m|T)$ applies (2 years in this case). More general formulations allowing these periods to be of different lengths are presented in Appendix A to show how to update the estimates of Equation 3 as more accident counts become available.

3.3/ Illustrating the Empirical Bayesian Method

To illustrate the method, data from LHRS Section # 28540 is used. This Class 2 road section is 0.6 km long, had an ADT of 5,900 and recorded 17 accidents in 1983-1984.

From Table 2 (later), the calibrated Level 1 (traffic volume only) model for a Class 2 road section is:

$$E(m|T) = (SCL) \times 1.2848 \times (ADT)^{0.9647} \qquad (5)$$

where the ADT is measured in thousands and, as mentioned above, E(m|T) is for the two year period 1983-1984.

From Equation 5, E(m|T) = 4.272 for a Class 2 Level 1 model. For this model, the estimated value of k for use in Equation 2 was 2.73 (also taken from Table 2), giving Var(m|T) = 6.685.

From Equation 3, E(m|T,x) = 12.037. This is the value of accident potential that should be used in the blackspot identification process and in estimating the effect of any treatment that might be applied. The variance of this estimate, Var(m|T,x), is 7.344 from Equation 4. Without a model, the best estimate of accident potential is 17 which, it appears, was a randomly high accident count.

More illustrative examples are provided in Appendix B.

3.4/ Further Exploration of the Methodology

Before proceeding to present and discuss the results, it is instructive to further explore the empirical Bayesian estimation procedure. To do so, it is necessary to rewrite Equation 3 as:

$$E(m|T,x) = wE(m|T) + (1-w)x \qquad (6)$$

$$\text{where } w = [1 + Var(m|T)/E(m|T)]^{-1}$$

Seen in this light, E(m|T,x), the estimated accident potential of a section is a mixture of what is observed, x, and of E(m|T) -- what is predicted on the basis of its characteristics (AADT, etc.).

The weight, w, of E(m|T) is always a number between 0 and 1. If the prediction has a large variance, i.e., there is much unexplained variation, then Var(m|T) >> E(m|T), w would be small and the accident potential would be close to x. In effect, little is learned from the measured characteristics of the section. Conversely, when Var(m|T) << E(m|T), the accident potential would be close to E(m|T) and little weight attaches to the observed count, x, which is given to random fluctuations.

Having recognized the importance of the value of w, it remains to explore what affects its value. This can be done by combining Equations 2 and 6 to give:

$$w = [1 + E(m|T)/k]^{-1} \qquad (7)$$

First, note from this equation that the larger the value of k, the larger is the value of w. In effect the size of k can be used as a relative measure of the variation explained by the regression model. Second, note that as E(m|T) increases, w decreases. From Equation 1, it can be seen that E(m|T) increases with section length and traffic volume. Thus, for a combination of long section lengths and/or high traffic volume, w can be small. In effect, for such sections, the accident potential would be closer to x, which is weighed heavily. This only makes sense, since the accident count for such sections is likely to be high and the random fluctuation relatively small.

To illustrate how the size of w is affected, in the above example for LHRS Section 28540, the value of w is 0.39; for a section half as long with the same traffic volume, w increases to 0.56. On the other hand, if Section 28540 was twice as long, w would be 0.24; if the section was 10 times as long, w would be 0.06.

This is an important result, since it means that, for sections that are long and/or have high traffic volumes, the accident count is a good estimate of accident potential, while for lower volume, shorter sections, the converse is true and the Bayesian estimates are definitely preferred. It is easy to see by extension that the same could be said if one substituted observation period for section length in the above statement. The conclusion would be that the longer the accident history, the more reliance can be placed on the accident count as an estimate of accident potential.

4/ Analysis and Results

This Chapter is divided into 2 sections. First the development of regression models of accident potential is described. This is followed by the empirical Bayesian estimation of accident potential for selected road sections used in a test on validity of the procedure.

4.1/ Regression Models for Estimating Accident Potential

Two levels of independent variables were used in regression models attempted for each road class identified in Table 1.

Level 1: Traffic volume
Level 2: Traffic volume, Geometric Variables or surrogates

Equation 1 of Chapter 3 indicates that the models are of the form:

$$E(m|T) = SCL*a1*(ADT)^{b1}$$

where, $Var(m|T) = E(m|T)^2/k$

This form ensures that predicted accidents would be zero for an ADT of zero, but does not, a priori, assume a linear relationship between accidents and traffic volume. A survey of relevant research by Satterthwaite (6) found that this model form is quite common.

The actual models estimated by GLIM were the log (base e) form:

$$\log[E(m|T)] = \log(SCL) + \log(a1) + b1*\log(ADT) \qquad (8)$$

In GLIM, the term log(SCL) is specified as an offset that is subtracted from each point estimate of log[E(m|T)].

Level 1 and Level 2 models are described separately below.

Level 1 Models:

The estimates of log(a1), b1 and k produced by GLIM are listed in Table 2 along with other relevant details.

Table 2/ Level 1 regression models

Model Parameter	Estimated Coefficient and (Standard Error)		
	CLASS 1 (Freeways)	CLASS 2 (Primary)	CLASS 3 (Secondary)
log(a1) for	-0.4656	0.2506	0.2605
ADT in thousands	(0.1020)	(0.02815)	(0.1166)
b1	1.024	0.9647	0.6923
	(0.03064)	(0.01995)	(0.07986)
Other parameters			
k	2.95	2.73	2.57
Scaled Deviance*	442	2003	206
Observations	404	1680	166

* *In Glim the statistic, scaled deviance, is an indicator of goodness of fit. In particular the significance of adding extra variables can be assessed by examining the reduction in scaled deviance.*

The estimated coefficients are for a 1 km section. For a section of length SCL km, the Class 2 regression model, for example, the 2 year accident potential would be estimated as:

$$E(m|T) = SCL * e^{0.2506} * ADT^{0.9647}$$

or,

$$E(m|T) = SCL * 1.2848 * ADT^{0.9647}$$

The fitted models seem to make sense intuitively. The ADT exponents (b1) indicate an almost linear relationship between accident occurrence and ADT for Classes 1 and 2, while for the lowest class (3), a decreasing effect of ADT on accident potential is indicated. One might speculate the ADT exponent for Class 3 roads reflects the fact that accident occurrence might be related to speed, which, in turn, might be strongly influenced by traffic volume for this class of road.

Before moving on, it should be pointed out that other possible traffic volume models -- incorporating the variable % commercial vehicles -- were attempted without success. The reduction in scaled deviance and increase in the value of k -- two measures of the importance of a new variable -- was negligible.

Level 2 Models:

The selected geometric variables varied with class of road, so each class is described separately.

Class 1:

Since freeway sections typically tend to be all multi-lane and divided, to have the same high geometric standards, and to be similar in other features like speed limit, it seemed unlikely that a model could incorporate geometric variables and be significantly better than the traffic volume model for this class. Indeed, it turned out that all attempts to add geometric variables, most notably, the number of lanes, were unsuccessful. Thus, there is no recommended Level 2 model for this class.

Class 2:

The class 2 road sample can be seen as a mixture of urban and rural roads, 2 and multi-lane roads and divided and undivided roadways, giving 6 possible categories (not 8, since there are no 2 lane divided highways). Pertinent details of the road sections in each category are shown in Table 3.

Table 3/ Categories for class 2 roads

	Rural			Urban		
	2-lane	Multilane		2-lane	Multilane	
		Divided	Undivided		Divided	Undivided
Sections	1400	49	90	73	8	60
Length, km	11413.1	188.7	219.9	106.3	5.3	91.3
Average ADT						
1983	2175	8459	11379	4683	16551	12396
1984	2218	8575	11886	4826	18074	13146
1985	2206	8631	12371	4925	19923	13435
1986	2337	9240	13152	5185	20891	14206
Accidents						
1983	12273	570	1429	369	44	1168
1984	13000	724	1631	370	53	1185
1985	13879	740	1660	406	74	1370
1986	12735	653	1767	400	102	1375

* *The urban category includes all roads with MTO road environment codes 2 and 3 (semi-urban and urban).*

Speeds and speed limit both vary significantly from category to category, but within each category the variation is quite small. Similarly, for each geometric characteristic that might affect accident occurrence, the variation within individual categories is small. Thus, it stands to reason that the use of a road section's category (as defined above) as a variable in an accident prediction model could be sufficient to account for variation due to geometric and speed factors. The final models for Class 2 roads reflects this reasoning. All attempts at incorporating other variables failed.

The final model form used the full Class 2 dataset and allowed the ADT coefficient (b1) to vary with the two categorical variables 2-lane/multi-lane and rural/urban while the coefficient a1 also varied with the categorical variable divided/undivided. A similar modelling procedure was used successfully by Maycock and Hall (7) who also used GLIM for their regression analysis.

GLIM provided estimates of log(a1) and b1 in Equation 7 for a base category (rural, undivided, 2-lane) and adjustments to be applied to these values for the other 5 categories. It turns out that the estimated coefficients obtained in this way are the same as those that would have been obtained by separately estimating Level 1 traffic volume models for each category. The estimated values of

log(a1) and b1 for the base category, along with the adjustments and the adjusted values for the other categories are shown in Table 4.

Table 4/ Level 2 regression models for class 2 roads

Model Parameter ADT in thousands	Adjustment(s) to base value	Standard Error	Adjusted Value
log(a1) for ADT in thousands for:			
Rural/undivided/2-lane	**0 (Base)**	**0.02783***	**0.2921**
Rural/undivided/multilane	-0.7186	0.2958	-0.4265
Urban/undivided/2-lane	1.003	0.1803	1.2951
Urban/undivided/multilane	0.0583	0.4507	0.3504
Rural/divided/multilane	-0.7186	0.2958	
	-0.3520	0.1137	-0.7785
Urban/divided/multilane	0.0583	0.4507	
	-0.3520	0.1137	-0.0016
b1 for:			
Rural/2-lane	**0 (Base)**	**0.02389***	**0.8310**
Rural/multilane	0.4727	0.1295	1.3037
Urban/2-lane	-0.2722	0.1138	0.5588
Urban/multilane	0.0453	0.2012	0.8763
Other Parameters:			
k=2.90; Scaled deviance=1882; Observations=1680			

* *These numbers apply to the base value; other standard errors apply to the adjustments.*

Again, the estimated coefficients are for a 1 km section. Thus, the accident potential of a rural, undivided, multi-lane road section of length SCL km, for example, could be estimated from:

$$E(m|T) = SCL * e^{0.2921} * ADT^{0.8310}$$

or,

$$E(m|T) = SCL * 1.3392 * ADT^{0.8310}$$

As for the Level 1 models, there is a pattern in the estimated ADT exponents that can be said to be in accord with intuition. The ADT exponents (b1) in Table 4 indicate that the effect on accident potential of increasing ADT is less pronounced for 2-lane roads and for urban roads. Again, it might be speculated that this is because speed is more likely to decrease with ADT for these roads. Figure 2 is a plot of the regression lines indicated by the estimated coefficients in Table 4 for each Class 2 category.

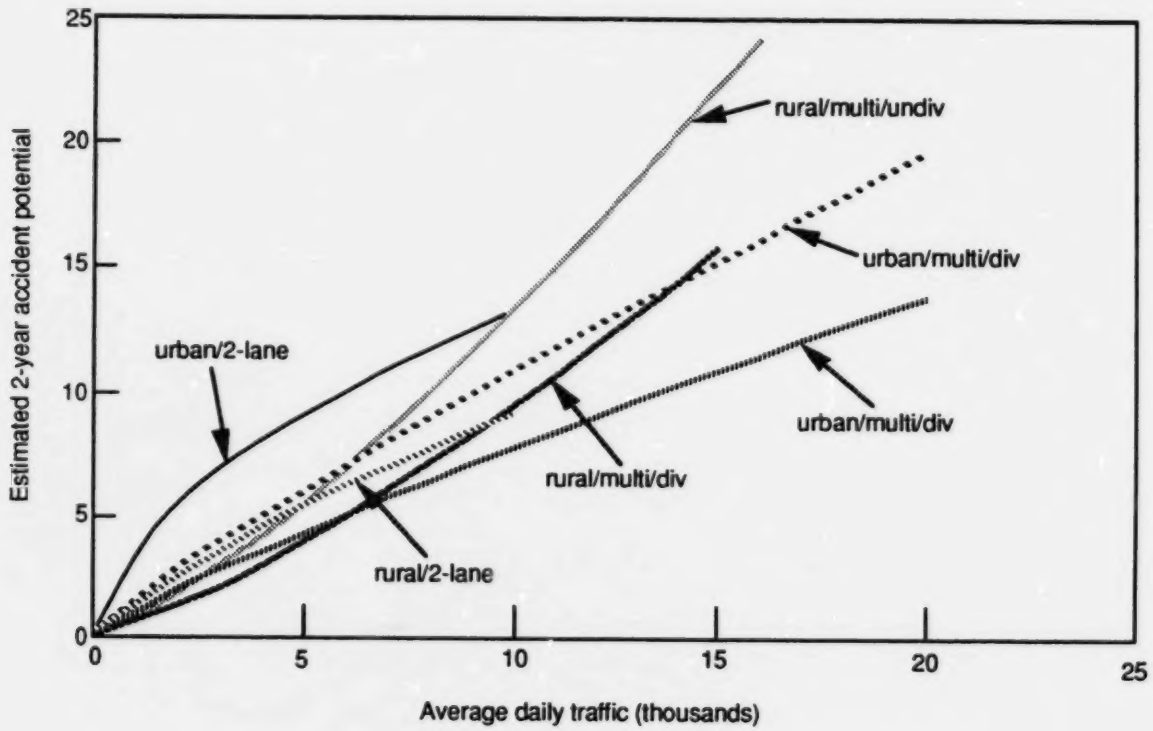

Figure 2/ Regression models for class 2 roads (1 km)

Class 3:

The modelling exercise for Class 3 roads was similar to that done for class 2 roads except that now the best explanatory variables turned out to be surface width and surface type. Additional explanatory variables provided little or no improvement particularly since the size of the dataset was small.

Exploratory analysis resulted in selection of surface width and surface type as categorical variables, each with 2 levels, giving a total of 4 category combinations. Table 5 describes these categories and provides relevant data for each category.

Table 5/ Categories for class 3 roads

	Pavement width ≤ 6.1 m		Pavement width > 6.1 m	
	High Class Surface*	Lower Class Surface*	High Class Surface	Lower Class Surface
Sections	11	75	55	25
Length, km	101.2	999.1	771.5	524.7
Average ADT				
1983	625	219	455	246
1984	607	220	477	247
1985	613	220	463	259
1986	637	229	484	271
Accidents				
1983	43	167	257	97
1984	34	186	259	122
1985	41	190	302	129
1986	43	162	248	100

* *High Class surfaces have MTO surface type codes >4 and consist of concrete, asphalt on concrete pavements and high class bituminous pavements. Lower class surface types have codes ≤ 4.*

The final model form used the full Class 3 dataset and allowed the ADT coefficient (b1) and the constant term log(a1) to vary with the two categorical variables narrow/wide pavement and low/high class surface. GLIM provided estimates of log(a1) and b1 in Equation 7 for a base category (pavement width ≤ 6.1 m, surface type ≤ 4) and adjustments to be applied to these values for the other 3 categories. The estimated values of log(a1) and b1 for the base category, along with the adjustments and the adjusted values for the other categories are shown in Table 6.

Table 6/ Level 2 regression models for class 3 roads

Model Parameter ADT in thousands	Adjustment to base value	Standard Error	Adjusted Value
log(a1) for ADT in thousands for:			
Narrow/low class	**0 (Base)**	**0.2477***	**0.003925**
Narrow/high class	-0.1028	0.5010	-0.1067
Wide/low class	0.2148	0.4425	0.2109
Wide/high class	0.2918	0.6394	0.2879
b1 for:			
Narrow/low class	**0 (Base)**	**0.1377***	**0.5844**
Narrow/high class	-0.1960	0.6268	0.3884
Wide/low class	0.1844	0.2809	0.7688
Wide/high class	0.0469	0.6878	0.6313
Other parameters:			
k=2.81; Scaled deviance=204.7; Observations=166			

* *Applies to the base value; other standard errors apply to the adjustments.*

As before, the estimated coefficients are for a 1 km section. Thus, the accident potential of a narrow Class 3 road section with a high class surface and length of SCL km, for example, could be estimated from:

$$E(m|T) = SCL * e^{-0.1067} * ADT^{0.3884}$$

or,

$$E(m|T) = SCL * 0.8988 * ADT^{0.3884}$$

Once again, it appears possible to rationalize the ADT exponents. The ADT exponent (b1) in Table 6 is lower for a pavement that is narrow than it is for a wider pavement. This pattern appears to be consistent with earlier speculation that the effect of increasing ADT on speed is more pronounced for roads that offer a lower quality of service. That the ADT exponent for a high class surface is lower than that for a low class surface does not at first glance appear to conform with this reasoning. However, examination of Table 5 reveals that the lower class surfaces tend to have lower volumes -- possible lower than the volume range for which speed might be affected. Figure

3 is a plot of the regression lines indicated by the estimated coefficients in Table 6 for each Class 3 category.

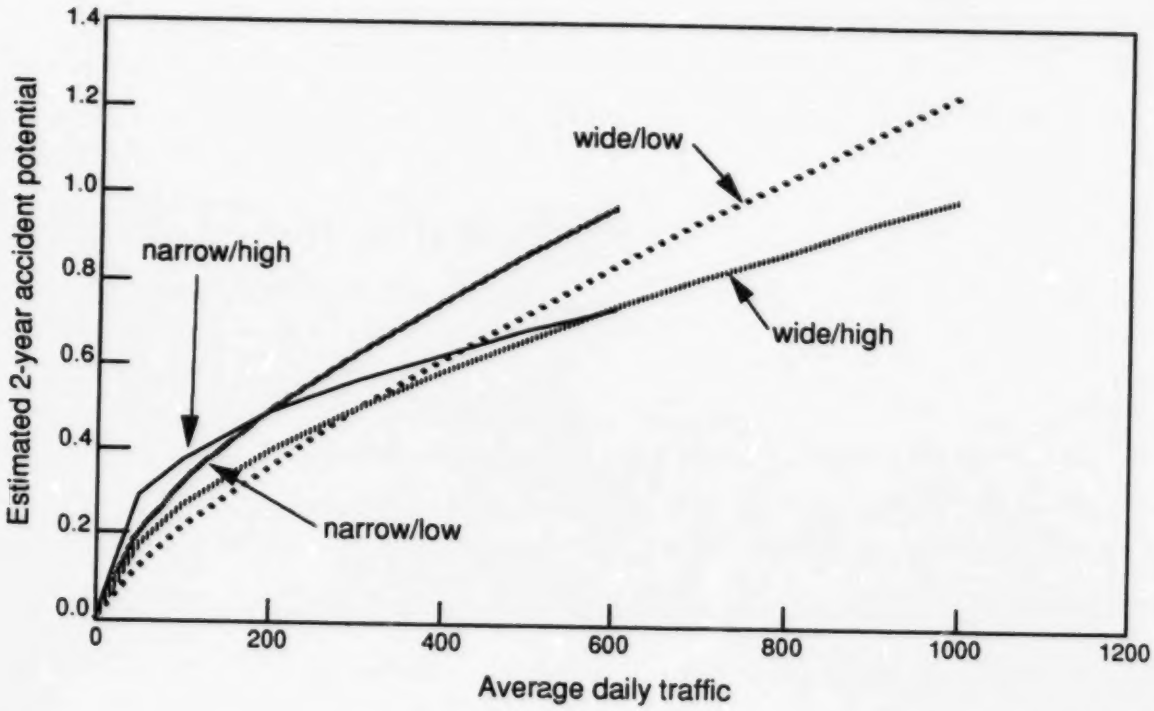

Figure 3/ Regression models for class 3 roads (1 km)

4.2/ Empirical Bayesian Estimation and Validation

In Chapter 3 is was shown how an initial accident potential estimate from regression models in Section 4.1 could be combined with the actual accident count on a road section to yield an empirical Bayesian estimate of accident potential for that section.

The point was made, however, that the need for this estimation process (as opposed to using the accident count as an estimate of accident potential) is more vital for sections with lower values of E(m|T) -- the regression model prediction. Since the data obtained was at the level of LHRS sections, most of which tend to be relatively long and/or to have high values of E(m|T), there could only be limited testing of the empirical Bayesian procedure with the current dataset.

Nevertheless, it was possible to demonstrate the usefulness of the models, using a sample of Class 2 roads. The sample consists of all 505 rural, 2-lane road sections in the dataset that have a length of 4 km or less. From this sample, the road section having the highest value of accidents/km in 1983-1984 in each of several ADT ranges was identified as a potential blackspot. (The width of the range varied to allow for 10-15 sections in each range.) From Equation 6 it can be seen that, since ADT is the now the only regression variable in this sample, the same sections would be identified if one were to properly base the selection on E(m|T,x) per km.

The 39 sections identified are listed in Table 7 along with other items, including the ADT, accidents, regression model predictions and empirical Bayesian estimates of accident prediction. Period 1 is 1983-1984, while period 2 is 1985-1986.

To illustrate how the estimates in Table 7 are derived, the calculations for the first row are shown. From Section 4.1, Table 4, the regression model prediction for ADT in thousands is:

$$E(m|T) = SCL*1.3392*ADT^{0.8310}$$

Thus, for LHRS section 26420,

$$E(m|T) = 1.2*1.3392*0.195^{0.8310} = \mathbf{0.413.}$$

From Equation 2, $Var(m|T) = [E(m|T)]^2/k$, where k=2.9 from Table 4. Thus,

$$Var(m|T) = 0.413^2/2.9 = \mathbf{0.059.}$$

Table 7/ Accident potential of selected rural 2-lane class 2 roads

Data for Periods 1, 2										
Section (km)			AADT		Accidents		E(m|T)	Var(m|T)	E(m|T, x)	Var(m|T, x)
LHRS#	Start	Length	1	2	1	2				
10375	0.0	1.3	1375	1475	3	0	2.268	1.774	2.590	1.137
25483	0.0	2.9	1400	1550	8	7	5.137	9.099	6.967	4.453
26810	0.0	2.1	1625	1600	33	24	4.210	6.112	21.258	12.587
21600	1.1	1.1	1675	1750	61	33	2.262	1.764	27.998	12.267
43800	0.0	3.5	1725	2150	15	8	7.374	18.750	12.847	9.221
17400	0.0	1.0	1875	1950	5	4	2.258	1.758	3.458	1.514
11900	0.0	1.6	2000	2000	8	8	3.812	5.010	6.190	3.516
16940	4.4	2.2	2200	2100	11	8	5.673	11.099	9.198	6.087
10360	0.0	1.1	2350	2450	7	4	2.996	3.096	5.031	2.557
11330	0.0	0.6	2500	2550	4	4	1.721	1.021	2.569	0.957
46250	0.0	2.2	2600	2625	16	12	6.518	14.650	13.080	9.053
10350	7.2	1.2	2750	2925	8	6	3.725	4.785	6.129	3.446
32410	5.4	2.0	2800	2925	9	8	6.302	13.695	8.150	5.581
20040	0.0	1.0	2950	3100	10	13	3.291	3.734	6.857	3.645
19410	0.0	0.7	3050	3325	8	9	2.368	1.934	4.900	2.203
14946	0.0	2.1	3225	3775	12	9	7.442	19.096	10.722	7.715
10770	8.4	2.1	3425	3550	12	12	7.823	21.104	10.870	7.931
20715	0.0	2.0	3625	2800	20	12	7.810	21.035	16.699	12.178
14680	0.0	1.2	3700	3725	10	5	4.767	7.835	8.020	4.987
34735	2.7	1.2	3850	6350	10	7	4.927	8.370	8.120	5.111
35240	6.3	1.1	4125	3975	11	9	4.783	7.887	8.653	5.387
23460	0.0	1.4	4225	3575	12	6	6.209	13.295	10.157	6.923
16140	0.0	1.6	4475	5450	11	11	7.444	19.106	10.003	7.198
19435	0.0	2.3	4600	4800	28	21	10.948	41.331	24.429	19.313
23706	4.7	1.9	4900	4550	24	15	9.532	31.328	20.625	15.814
12200	3.9	4.0	5250	5400	37	44	21.251	155.720	35.109	30.893
29500	1.9	0.4	5600	5850	6	0	2.242	1.734	3.881	1.692
28540	0.0	0.6	5900	6050	17	12	3.512	4.254	10.900	5.970
33010	1.8	0.8	6050	6300	8	7	4.782	7.885	6.785	4.224
40330	2.8	1.6	6850	6650	34	27	10.603	38.768	28.975	22.752
35430	0.0	1.5	7525	7150	23	23	10.748	39.834	20.397	16.063
16490	1.9	1.4	8500	8950	26	33	11.100	42.488	22.914	18.167
16100	0.0	2.2	10200	10400	49	44	20.297	142.055	45.412	39.734
10550	0.0	0.8	12450	13650	38	30	8.710	26.162	30.684	23.020
Totals		**60.5**			**641**	**503**	**230.9**	**750.9**	**487.1**	**338.3**
(Adjusted)*					**(675)**				**(513)**	**(375)**

* *Adjustments account for a 5.3% increase in accidents from period 1 to 2.*

From Equation 6, $E(m|T,x) = wE(m|T) + (1-w)x$, where x is the period 1 accident count and $w = 1/[1 + E(m|T)/k]$ from Equation 7. (Note that Equation could have been used instead.)

Thus, for Section 26420, $w = 1/[1+0.413/2.9] = 0.875$, and,

$$E(m|T,x) = 0.8753*0.413 + (1-0.8753)*8 = \mathbf{1.359}$$

Finally, to estimate $Var(m|T,x)$, the above values of $E(m|T)$ and $Var(m|T)$ can be used in Equation 4 to give:

$$Var(m|T,x) = \mathbf{0.169}.$$

Several observations follow from the data in Table 7. First, note that there is a substantial decrease in accident count from period 1 to period 2. Assuming that the sections were largely untreated during the 4 year period (a reasonable assumption according to information in the inventory file), the logical conclusion is that this difference occurs because of a random up-fluctuation in the accident count during period 1. The difference is even more pronounced if the period 1 count is adjusted to account for the fact that rural Class 2 roads had a 5.3% increase in accidents in period 2.

The second observation is that, since the sections were not selected on the basis of the period 2 accident count, this count is an unbiased estimate of true accident potential of these sections. Thus, an unbiased estimate of the total of the true accident potentials of these sections is 503. Thus, using the period 1 accidents as estimates of accident potential overestimates this value by 34% !

The third, and most important, observation is the close correspondence between the total of the Bayesian estimates, $E(m|T,x)$ (487, or, the adjusted value of 512), and the total of the estimates of true accident potential (503). This correspondence serves to validate the Empirical Bayesian procedure.

Finally, note that even though the period 2 accident count is an unbiased estimate of the true accident potential, the variance of this estimator (equal to the count, assuming a Poisson distribution of counts on a section) is substantially higher than that of the Empirical Bayesian estimator (shown in the last column of Table 7).

Though the final observation is of considerable interest, it is immaterial when considering the practical need to use present accident counts to estimate **future** accident potential of potential blackspots. In such an application, the only alternatives estimators are the recent accident count and the Empirical Bayesian estimator. On the basis of the validation exercise just described, it is clear which is preferred.

5/ Summary, Conclusions and Recommendations

The research used data available at the level of total accidents on relatively long LHRS sections. In spite of this limitation it can be said that useful insights were gained into the estimation of accident potential for Ontario road sections. Following is a summary and important conclusions of this work, followed by a set of recommendations.

5.1/ Summary and Conclusions

1. Theory and empirical results demonstrated that accident counts are, in general, not good estimators of accident potential, particularly for short road sections and those that might be of interest when considering and applying treatment in a blackspot identification and treatment process.

2. The accident count is a reasonable estimator of accident potential only if sections are identified at random, i.e., without regard to the accident count itself, or for sections with a high count and relatively small random fluctuation.

3. An empirical Bayesian procedure was proposed for the unbiased estimation of accident potential. This estimator combines the accident count information with a regression model prediction of accidents expected on the basis of the section's traffic volume and geometric characteristics.

4. Both the Level 1 and Level 2 regression models seemed to make sense intuitively. However, these regression models were intended for use as accident predictors and not as an exploration of the causal factors related to accident occurrence. Thus, they should not be used to answer what/if questions on the impact of altering values of the independent variables.

5. When the regression predictions were combined with the accident count information for a set of potential blackspots the resulting empirical Bayesian estimates were seen to be not only unbiased but also to have a smaller variance, even when the accident count estimator is unbiased.

5.2/ Recommendations

1. The empirical Bayesian procedure should be used for estimating accident potential of Ontario road sections. The models developed in this work could be used if one makes the reasonable assumption that the 1983-1984 relationship between accident occurrence and the independent variables is still valid. Suggested uses of the procedure developed are listed below:

a) **In estimating the safety effect of improvements that might have affected safety.** The point estimate of treatment effect is equal to the % change in accident potential before and after treatment. The empirical Bayesian estimate should be used for the accident potential before treatment. The accident count is at present the most reliable estimate of accident potential after treatment. The estimate of safety effect for a single project is likely to have a high variance that can be reduced if estimates for several projects are combined. This issue is explored in other sources (7).

b) **In identifying potential blackspots.** With the current dataset it was not possible to explore alternative blackspot identification processes in any depth. Yet, some suggestions might be in order. Since treatment type, cost and policy considerations might vary according to the such factors as class, environment and traffic volume of a road section, one might consider a procedure similar to the one attempted in this research. Road sections in each of a number of class, environment and volume categories could be identified for further analysis on the basis of their accident potential per km.

c) **In estimating the safety benefits (or disbenefits) of potential treatments at blackspots or other sites being considered in a resource allocation procedure.** To do so, the best estimates of safety effect of a treatment should be applied to the current value of accident potential of a site being considered for the treatment. An example application could be in deciding which of the sites being considered for installation of climbing or passing lanes should be given the highest priority.

d) **In updating estimates of accident potential with new accident information using methods outlined in Appendix A.**

2. MTO should seek to build improved regression models of accident potential. To do so the accident data needs to be available at a finer level such as the details of each accident on a road section. Separate models could then be constructed for single vehicle accidents and collisions, for intersection and non-intersection accidents. More importantly, the modelling effort could attempt to relate accident occurrence to the specific traffic volumes at the time of accidents.

3. The model building should be extended to develop models for estimating the accident potential of intersections because of the disproportionate numbers of accidents that occur at these locations. Also, similar models should be constructed for roads and intersections under the jurisdiction of municipalities.

4. With the availability of more complete accident data and improved accident potential models, the exploration of alternative blackspot identification processes should be pursued.

References

[1] Taylor, J.I., and H.T. Thompson, "Identification of Hazardous Locations", Federal Highway Administration report FHWA-RD-77-83, 1977.

[2] Farr, E.H., "Rail-Highway Crossing Resource Allocation Model", Federal Railroad Administration report FRA-RRS-001, 1981.

[3] Hauer, E., and B.N. Persaud, "How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effect of Warning Devices", Transportation Research Record 1114, pp. 131-140, 1987.

[4] Hauer, E., Smiley, A., and B.N. Persaud, "Accidents, Convictions and Demerit Points: An Ontario Driver Record Study", Ontario Ministry of Transportation report 491-01, 1988.

[5] Baker, R.J., and J.A. Nelder, "Generalized Linear Interactive Modelling. The GLIM System -- Release 3", Rothamsted Experimental Station, Harpenden, U.K., 1978.

[6] Satterthwaite S.P., "A Survey of Research into the Relationships Between Traffic Accidents and Traffic Volumes", Transport and Road Research Laboratory Supplementary report 692, U.K., 1981.

[7] Maycock, G., and R.D. Hall, "Accidents at 4-arm Roundabouts", Transport and Road Research Laboratory Laboratory report 1120, U.K., 1984.

[8] Hauer, E., Lovell, J., and B.N. Persaud, "New Directions for Learning about Safety Effectiveness", Federal Highway Administration report FHWA/RD-86-015, January 1986.

Appendix A: Updating Empirical Bayesian Estimates

The procedure for updating Empirical Bayesian estimates of accident potential with new accident count information is rather straightforward and follows from the equations developed earlier. Two cases are possible:

(i) the characteristics, e.g. traffic volume, of the section have changed.

(ii) the characteristics have not changed.

In case (i), present theory suggests that the accident potential estimates should not be updated since the characteristics have changed; rather, new values of accident potential are estimated using the latest year's values for the characteristics and the accident count.

For case (ii), the updating procedure is as follows.

Let n = the number of regression model periods for which the new accident counts are available. Thus, for example, for the 2-year models developed, n=1 if 2 years of accident counts become available; if 1 year of counts becomes available, n=0.5.

The updating of accident potential is accomplished by re-writing the main Bayesian equation (3) in a more general form to allow the period over which the accident count is taken to differ from the period for the regression model predictions.

It can be shown that the general forms for Equation 3 and 4 are:

$$E(m|T,x,n) = (x+q)/(n+p) \qquad (A1)$$

$$Var(m|T,x,n) = (x+q)/(n+p)^2 \qquad (A2)$$

where,

$$p = E(m|T)/Var(m|T)$$

$$q = [E(m|T)]^2/Var(m|T)$$

Suppose there is a current value of E(m|T,x,n) and that a further n periods of accident data become available; then new estimates of E(m|T,x,n) and Var(m|T,x,n) are obtained by simply substituting the current values in place of E(m|T) and Var(m|T) in the above equations. To illustrate, in the example in Section 3.2, the road section had estimates of E(m|T,x,1) = 12.037 and Var(m|T,x,1) = 7.344 for a 2 year period. If the road section had 7 accidents and the same ADT in the next year, then using x=7 and n=0.5 in Equations A1 and A2,

$$p = 12.037/7.344 = 1.639$$

$$q = 12.037^2/7.344 = 19.929.$$

The updated value of E(m|T,x,n) is:

$$(7 + 19.929)/(0.5+1.639) = 12.590$$

The variance of this estimate is $26.929/2.139^2 = 5.886$.

Both of these values still apply to a 2-year period. Notice that the variance is considerable reduced with the extra year's accident count information.

Figure A1 illustrates the progressive improvement in estimates of accident potential for the section as more information is utilized. The plotted lines are the gamma probability density functions for the three estimates of accident potential for the section. The initial estimate is the regression prediction for which the accident information for the section was not used. This estimate has the largest variance. The second estimate uses the accident count for a 2-year period in the empirical Bayesian method. The variance is considerably reduced. The final estimate comes from the above example and updates the second estimate using an additional year's accident count. This estimate has the least variance.

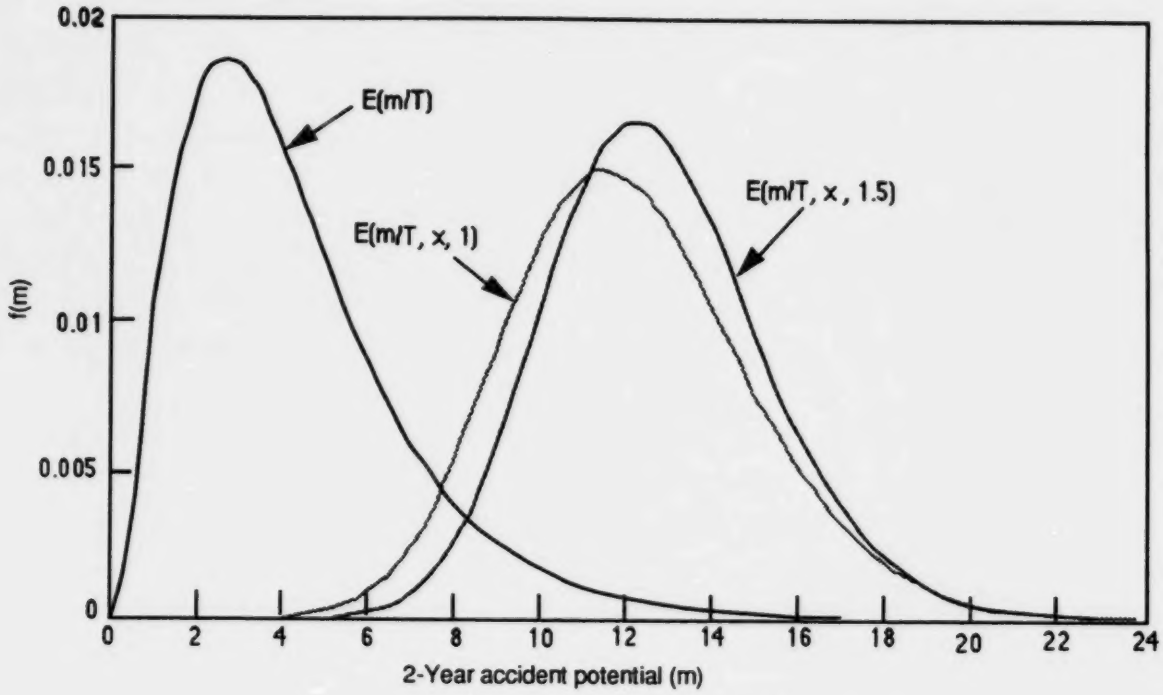

Figure A1/ Gamma distribution of accident potential

Appendix B: Illustrative Applications

Two detailed examples are presented below to illustrate how the procedures developed in this research might be applied. It is important to note that there are other potential applications.

Illustration 1: Selection of Sites for Upgrading

Consider a district with several hundred kilometers of Class 3 roads. It is of interest to consider a few short (say, 1 km) sections for low cost improvements. The question is which sections should one consider for detailed inspection. Assume that a practical number is 25. The procedures described in this report can be used to select the 25. Detailed inspection of the 25 will result in, say, 5 sections which can be treated under the current budget for low cost improvements to Class 3 roads.

The 25 sections could be selected by ranking all sections of interest on the basis of their accident potential and then selecting the top 25. Accident potential is estimated from the appropriate models in Chapter 4. For example, a 1 km Class 3 road section with a low class surface and a wide pavement which had 4 accidents and an ADT of 500 in the past 2 years would have (from Table 6):

$$E(m|T) = 1*e^{0.2109}*0.5^{0.7688} = 0.7247$$

$$Var(m|T) = 0.7247^2/2.81 = 0.1869$$

$$E(m|T,x) = 4 + [0.7247/(0.7247+0.1869)]*(0.7247 - 4) = 1.396$$

The estimate of accident potential for 2 years is 1.396.

Suppose another 1 km section being considered had 5 accidents in the past 2 years, a high class surface, a narrow pavement, and an ADT of 300. For this section,

$$E(m|T) = 1*e^{-0.1067}*0.3^{0.3884} = 0.5631$$

$$Var(m|T) = 0.5631^2/2.81 = 0.1128$$

$$E(m|T,x) = 5 + [0.5631/(0.5631+0.1128)]*(0.5631 - 5) = 1.304$$

Thus, in the ranking of sections, the first site would be ranked higher than the second even though it had fewer accidents in the past 2 years.

Illustration 2: Evaluation of Treatment Effectiveness

Imagine that the same treatment was applied at 5 sites which in the 2 year period before treatment had values of E(m|T,x) of 2.123, 4.245, 6.968, 5.402, and 3.576. In the after period, these sites recorded 2, 4, 3, 4, and 3 accidents respectively. Province-wide, similar untreated sites had a 4% increase in accidents between the 2 periods.

The accident potential estimates are used to estimate what would have happened collectively at the 5 sites in the after period had the treatment not been applied. This estimate is:

$$1.04*(2.123+4.245+6.968+5.402+3.576) = 23.21$$

With the treatment the total accident count was 2+4+3+4+3 = 16.

Thus the estimate of treatment effectiveness is:

$$100*(23.21-16)/23.21 = 31.1\%$$

Note that, since only 5 sites with relatively small counts were treated, this estimate of treatment effectiveness would be subject to large uncertainty. Methods for expressing this uncertainty and for updating it as the results of subsequent treatments become available are explored in (**8**).